Live–virtual–constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part 2: demonstration of the framework

Mansikka, H. P., Virtanen, K., Harris, D. & Salomaki, J.

DOI 10.1177/1548512919886378
ISSN 1548-5129
ESSN 1557-380X

Publisher: SAGE Publications

LIVE-VIRTUAL-CONSTRUCTIVE SIMULATION FRAMEWORK FOR TESTING AND EVALUATION OF AIR COMBAT TACTICS, TECHNIQUES AND PROCEDURES, PART 2: IMPLEMENTATION

Heikki Mansikka

Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland

Insta DefSec, Tampere, Finland

Kai Virtanen

Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland Department of Military Technology, Finnish National Defence University, Helsinki, Finland

Don Harris

Faculty of Engineering and Computing, Coventry University, Coventry, United Kingdom

Jaakko Salomäki

Department of Military Technology, Finnish National Defence University, Helsinki, Finland

ABSTRACT

This paper demonstrates the implementation of the L-V-C simulation framework discussed in Part 1 of this study. The implementation is demonstrated in the testing and evaluation (T&E) of air combat TTP. TTP consists of rules that describe how aircraft pilots coordinate their actions to achieve goals in air combat. In the demonstration, TTP rules are developed iteratively in separate C-, V- and L-simulation stages. In C-stage, the optimal rules with respect to probabilities of survival (P_s) and kill (P_k) of aircraft are determined without considering the impact of human behavior in human-machine interaction (HMI). In V-stage, the optimal rules are modified by assessing their applicability with P_k and P_s, as well as HMI measures regarding pilots' situation awareness, mental workload and TTP rule adherence. In L-stage, F/A-18 aircraft and qualified fighter pilots are used to evaluate whether the TTP developed in C- and L-stages leads to acceptable P_k, P_s, and HMI measures in a real-life environment While this paper concentrates on air combat, the principles of this demonstration can be applied to any military or civil simulation study where HMI is of concern.

Keywords: air combat, human factors, human-machine interaction, live-virtual-constructive, mental workload, performance, simulation, situation awareness, testing and evaluation

1. INTRODUCTION

This paper demonstrates how the L-V-C simulation framework discussed in Part 1 of this study is implemented in the testing and evaluation (T&E) of air combat tactics, techniques and procedures (TTPs). First, the basic principle of the L-V-C simulation framework is briefly summarized. Any reference to Part 1 denotes a reference to Mansikka et al.[1] For a full description of the theoretical framework, the readers are referred to Part 1.

The L-V-C framework utilizes live- (L), virtual- (V) and constructive simulations in TTP T&E. TTP consists of rules that describe how aircraft pilots coordinate their actions to achieve goals in air combat. In the demonstration, the L-V-C simulation framework is implemented to identify the operationally HMP optimal values (see Part 1) for the selected rules of wingmen. Wingmen are members of a flight, which is a unit of four aircraft. A flight is composed of two 'elements', a lead element and a wing element. The elements have two aircraft in each, the leader and the wingman.

When the L-V-C simulation framework is used for the TTP T&E, a scenario along with the initial TTP and its associated TTP rules are first defined. The scenario describes the friendly and enemy aircraft involved, and their primary goals and the TTP rules describe how the friendly aircraft can best achieve their goals in the given scenario.

As illustrated in Figure 1, the L-V-C simulation framework has C-, V- and L-stages. In the first C-stage, the quantitative rules of the initial TTP are implemented into C-simulation and the enemy aircraft are set to follow the behaviors determined in the initial scenario. C-simulation runs are conducted until machine performance (MP) optimal values (see Part 1) maximizing the optimization criterion probability of kill (P_k) and fulfilling the constraint probability of survival (P_s)=1 are found. Wingmen's situation awareness (SA), mental workload (MWL), normative performance (NP) or human-machine performance (HMP) output are not considered in C-stage. For a full description of SA, MWL, NP, HMP and their measurement, see Part 1. If a C-stage is repeated due to an unacceptable SA, MWL, NP or HMP output in V- or L-stages, (see the dashed lines from V- and L-stages to C-stage in Figure 1), the original optimization criterion is relaxed by the minimization of $(P_k\text{-}P_{kref})^2$, where the reference probability of kill, denoted by P_{kref}, is selected based on the results of V- or L-stage as well as on the optimal values of P_k obtained in earlier C-stages. By analyzing the results of the

preceding V- or L-stage, the quantitative rules whose values should be adjusted in the new C-stage are selected.

[insert Figure 1]

The first V-stage considers both the qualitative rules of the initial TTP and the MP optimal quantitative rules originating from C-stage. If V-stage is repeated, the qualitative rules originate from the preceding V- or L-stage (see the dashed line from L-stage to V-stage, and the dotted line from V-stage to V-stage in Figure 1).

The wingmen fly the V-simulation as participants, while all other aircraft are implemented in the simulation as constructive entities. Wingmen's NP, SA, MWL, and HMP output are recorded. HMP output is measured using P_k and P_s. The estimation of P_k is based on the ratio of enemy aircraft alive at the beginning and at the end of the simulation, and the estimation of P_s is based a ratio of friendly aircraft alive at the beginning and at the end of the simulation. The friendly constructive entities are set to follow the rules derived in the preceding stages of TTP T&E and the enemy constructive entities are set to follow the same scenario as in C-stage. The participants are tasked to follow the directed qualitative rules and MP optimal quantitative rules. Participants are not told how the scenario unfolds.

If P_k and P_s are unsatisfactory at the completion of V-stage, the rules that could be revised to improve the overall HMP output are identified. If, however, P_k and P_s are satisfactory, the objective is to identify the rules that could improve NP, SA or MWL. If the quantitative rules are modified, TTP is returned to C-stage without modifying the qualitative rules (see the dashed line from V- to C-stage in Figure 1). If the qualitative rules are modified, V-stage is repeated with refined verbal descriptions of the participants' qualitative rules (see the dotted line from V- to V-stage in Figure 1). The constructive entities' qualitative rules are adjusted only if they affect the participants' ability to adhere to their rules.

Each time V-stage is repeated, NP, SA, MWL, P_k and P_s are compared to those of the preceding V-stage, with the aim of identifying any significant differences between the simulations' P_k, P_s, and NP, SA, and MWL scores. Once the outcome of V-stage is satisfactory, HMP optimal rules in the simulated environment are obtained and the TTP T&E progresses to L-stage.

In L-stage, the previously determined HMP optimal rules are evaluated in a real-life environment. Real aircraft and pilots are used in L-stage. The participants are tasked to follow the HMP optimal rules defined in V-stage. All other pilots serve as supporting pilots and follow the constructive entities' rules used in the preceding stage. HMP output, measured by P_k and P_s, and the participants' scores of NP, SA and MWL are recorded. Here, P_k and P_s are estimated in a same way as in the V-simulations. The results of V- and L-stages are comprehensively compared. The results are balanced if L-stage's P_k and P_s are acceptable and the scores of NP, MWL and SA are not significantly worse than those obtained at V-stage. If this is not the case, the potential rules for revision are identified in a same fashion as at V-stage. Then, TTP is returned to C- or V-stage depending on the need for either qualitative or quantitative rule adjustments (see the dashed lines from L-stage to C- and V-stages in Figure 1). If V- and L-stages' results are balanced, TTP T&E is complete. The implementation of the L-V-C simulation framework is demonstrated in the following section.

2. DEMONSTRATION OF THE L-V-C SIMULATION FRAMEWORK

2.1 Initial TTP

A flight's initial TTP with quantitative rule values and qualitative rule descriptions for a beyond-visual-range (BVR) defensive counter air scenario was defined. The scenario had three seamlessly connected engagements against a numerically superior enemy. An engagement refers to an isolated attack against a threat with a directive or authorisation to use sensors and/or weapon systems against a designated target[2]. Each engagement had the following phases: 1) target assignment, search, and identification, 2) weapon employment, and 3) evasion and egress. The engagements' complexity and SA demands were designed to increase towards the third engagement.

The scenario started with a long-range BVR engagement, where the flight's goal was to employ weapons to designated targets while maximising range to the enemy aircraft. Once the flight had reached the evasion and egress phase of the first engagement, it initiated a short-range engagement against the surviving enemy aircraft. After the second engagement, the flight executed a third engagement against the remaining enemy aircraft, which were now chasing the flight. The air combat task ended when the flight was in the evasion and egress phase after the third engagement. During the scenario, both wingmen were tasked to launch

three missiles, one in each engagement. The enemy aircraft represented modern air-superiority fighters, which used predefined TTP to kill all friendly aircraft. Figure 2 outlines the scenario used in the demonstration.

[insert Figure 2]

Seventy-six rules from the initial TTP were selected for the NP measure. While the rule values and descriptions contain classified information, Table 1 summarizes the TTP rules without their values. The L-V-C simulation framework was used to identify the wingmen's operationally HMP optimal quantitative rules related to missile launch ranges (rules 7, 29 and 52), evasive maneuver ranges (rules 14, 36 and 59), and the durations of the egress phases (rules 17, 40 and 62). In addition, the optimal verbal descriptions of the qualitative rules presented in Table 1 were recognized.

Table 1. TTP rules. The quantitative rules designated with an asterix (*) were adjusted in C-stage. The qualitative rules designated with a double asterix (**) were refined in V-stage.

Engagement			TTP rule
1	2	3	
Rule	Rule	Rule	
1			Airspeed from 30 seconds to 1 minute since the simulation start
2			Altitude from 30 seconds to 1 minute since the simulation start
	23		Airspeed from 4 minutes to 4 minutes 45 seconds since the simulation start
	24		Altitude from 4 minutes to 4 minutes 45 seconds since the simulation start
		46	Airspeed from 5 minutes to 5 minutes 45 seconds since the simulation start
		47	Altitude from 5 minutes to 5 minutes 45 seconds since the simulation start
3	25	48	Radar search parameters
4**	26**	49**	Target's declaration and identification
5	27	50	Pre-missile launch maneuvering
6	28	51	Target engagement decision
7*	29*	52*	Missile launch range
8	30	53	Missile launch parameters
9**	31**	54**	Communication of the missile's inflight phase changes
10	32	55	Angle of the post-launch maneuver
11	33	56	Timing of the missile's data link termination
12	34	57	Timing of the electronic countermeasures activation

13	35	58	Post-missile launch maneuvering
14*	36*	59*	Range of the evasive maneuver initiation
15	37	60	Missile's target at end-game
16	39	61	G-load during the evasive maneuver
	38		Duration of the doppler notch maneuver
17*	40*	62*	Egress phase duration
18	41	63	Egress phase heading
19	42	64	Range between flight members
20	43	65	Level of mutual support between flight members
21**	44**	66**	Communication of tactical status
22	45	67	Deconfliction to flight members, terrain and obstacles

2.2 First C-stage (C1)

The quantitative rules for the initial TTP and the flow of the enemy aircraft were implemented into a constructive Air Combat Evaluation Model (ACEM) simulation. ACEM is a Raytheon built air combat simulation, typically used for studying operational-level requirements, preliminary designs and tactical utility of TTPs at the engagement level. ACEM has been widely applied for conducting evaluation and test studies of aircraft and systems used in this demonstration, and therefore its validity can be considered adequate.

The wingmen's quantitative rules were adjusted until MP optimal values were found for the missile launch ranges (rules 7, 29 and 52), the evasive maneuver ranges (rules 14, 36 and 59), and the egress phase durations (rules 17, 40 and 62). MP optimal values resulted in P_s=1.0, while P_k for the first missile launch was 0.77. The second and third missile launches resulted in P_k of 0.72 and 0.71, respectively. MP output, measured as the average of P_k values, was 0.73.

2.3 First V-stage (V1)

A Boeing built Weapon Tactics and Situation Awareness Trainer (WTSAT) was used to run the V-simulations. WTSAT is an operationally used, non-motion F/A-18C flight simulator, with a 135-degree field of view and a fully functional cockpit. WTSAT replicates the F/A-18C flying

characteristics, weapons, and aircraft systems with such an accuracy that pilots can use it to fly their annual proficiency check flights. Fourteen combat ready male F/A-18C pilots were recruited as participants. Their average flying experience in the F/A-18C was 737 flight hours (SD=352). All participants were fit to fly and were qualified to fly the scenario and the associated TTP.

The participants flew the mission as wingmen while all other aircraft were constructive entities. The constructive element leaders were programmed and the participants were tasked to adhere to MP optimal rules defined in C1 and the initial TTP's qualitative rules. The enemy aircraft were designed to follow the scenario defined in the initial TTP. The participants' NP, SA and MWL scores were assessed. The overall duration of the flying task was 7 minutes 36 seconds.

2.3.1 V1 NP Results

The rules listed in Table 1 were used to calculate the NP score. Each rule adhered to was given a score of 1, whereas each rule omitted or not adhered to was given a score of 0. The mean score for each rule was used as its NP score. The overall mean NP score in V1 was 0.76 (SD=0.42). Table 2 summarizes the mean NP scores for each engagement. Friedman test revealed significant differences between the NP scores in engagement 1 ($\chi^2_{(21)}$=83.74, p<0.001), in engagement 2 ($\chi^2_{(22)}$=70.38, p<0.001), and in engagement 3 ($\chi^2_{(21)}$=71.32, p<0.001). Figure 3 presents the mean NP scores across the TTP rules in engagements 1-3.

Table 2. Means and standard deviations (SD) of the NP scores in each engagement (N=14).

	Mean	SD
Engagement 1	0.82	0.36
Engagement 2	0.76	0.41
Engagement 3	0.70	0.39

[insert Figure 3]

2.3.2 V1 SA Results

SA was measured using 45 probes, 15 in each engagement. Table 3 lists the SA probes used. Each correct answer to a probe was given a score of 1, whereas each incorrect answer was scored as 0. The mean of the probe's scores was used as its SA score. The overall mean SA score for V1 was 0.81 (SD=0.37). Table 4 summarizes the descriptive statistics of the SA and SA level scores in each engagement. Friedman test revealed significant differences in the SA scores in engagement 1 ($\chi^2_{(14)}$=42.16, p<0.001), engagement 2 ($\chi^2_{(14)}$=48.81, p<0.001), and in engagement 3 ($\chi^2_{(14)}$=38.77, p<0.001). Figure 4 displays the mean SA scores across the SA probes in engagements 1-3.

Table 3. SA probes, their descriptions and associated SA levels.

Engagement				
1 Probe	2 Probe	3 Probe	SA Level	Description
1	16	31	1	Did you correctly perceive your and the flight members' positions with respect to enemy?
2	17	32	1	Did you correctly perceive the enemy aircraft positions and geometries?
3	18	33	1	Did you correctly perceive the declarations and types of enemy aircraft?
4	19	34	1	Did you correctly perceive which enemy aircraft were targeted?
5	20	35	1	Did you correctly perceive the flight members' roles and duties during the engagement?
6	21	36	2	Did you correctly understand the flight members' positions with respect to timeline?
7	22	37	2	Did you correctly understand the directed TTP?
8	23	38	2	Did you correctly understand if the flight was adhered to the directed TTP?
9	24	39	2	Did you correctly understand the changes in the enemy presentation?
10	25	40	2	Did you correctly understand the changes in the flight members' tactical statuses?
11	26	41	3	Were you able to correctly anticipate how the engagement evolved?
12	27	42	3	Were you able to correctly anticipate the actions of your flight members?
13	28	43	3	Were you able to correctly anticipate the ranges and decision points in the timeline?

14	29	44	3	Were you able to generate alternative courses of action in case something unexpected would happen?
15	30	45	3	Did you correctly anticipate the result of the engagement?

Table 4. Means and standard deviations (SD) of the SA and SA level scores in each engagement (N=14).

	SA		SA level 1		SA level 2		SA level 3	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Engagement 1	0.94	0.24	0.94	0.23	0.99	0.12	0.89	0.32
Engagement 2	0.76	0.43	0.86	0.35	0.84	0.37	0.57	0.50
Engagement 3	0.73	0.45	0.87	0.34	0.77	0.42	0.54	0.50

[insert Figure 4]

2.3.3 V1 MWL Results

The score on the MWL measure was determined by the pilot's ratings on NASA Task Load Index (NASA-TLX[3]) scale's dimensions using a scale from 0 (low MWL) to 100 (high MWL). The overall mean of the MWL score in V1 was 47.50 (SD=14.44). Table 5 presents the descriptive statistics of the MWL dimensions. Friedman test revealed significant differences between the MWL dimensions ($\chi^2_{(5)}$=24.81, p<0.001). Based on Wilcoxon signed ranks test, the effort dimension differed significantly from the physical (*Z* =-2.23, p<0.05), performance (*Z* =-1.97, p<0.05), and frustration dimensions (*Z* =-2.59, p<0.05). The temporal dimension differed significantly from the physical (*Z* =-2.96, p <0.01), performance (*Z* =-1.96, p<0.05), and frustration dimensions (*Z* = -2.52, p<0.05). Furthermore, the mental dimension differed from the frustration dimension (*Z* =-2.13, p<0.05). All other pairwise comparisons were non-significant.

Table 5. Means and standard deviations (SD) of the MWL dimensions (N=14).

MWL dimension	Mean	SD
Mental	52.14	20.82
Physical	41.43	21.79
Temporal	60.00	18.81
Performance	44.29	20.65
Effort	56.43	25.90
Frustration	30.71	20.93

2.3.4 V1 Conclusions

In V1, P_k and P_s were both 1.0, which implies the satisfactory HMP output. In engagement 1, the lowest NP scores were associated with rules 12 (timing of the electronic countermeasures), 2 (altitude from 30 seconds to 1 minute since simulation start), and 7 (missile launch range) (see Figure 3). In engagement 2, the lowest NP score was provided by rule 38 (duration of the doppler notch maneuver) (see Figure 3). In engagement 3, rules 59 (range of the evasive maneuver initiation), 52 (missile launch range), 47 (altitude from 5 minutes to 5 minutes 45 seconds since the simulation start), and 53 (missile launch parameters) had the lowest NP scores (see Figure 3). In all engagements, the lowest SA scores were associated to the SA probes related to SA level 3, i.e., probes 14, 29 and 44 (were you able to generate alternative courses of action in case something unexpected would happen?), probe 27 (were you able to correctly anticipate the actions of your flight members?), and probe 43 (were you able to correctly anticipate the ranges and decision points in the timeline?) (see Table 4 and Figure 4). Overall, the temporal dimension of MWL was high (see Table 5).

Based on the analysis by SMEs, rule 7 (missile launch range in engagement 1) was identified as the quantitative candidate with the most potential for revision. The SMEs reasoned that by increasing the missile's launch range in engagement 1, the temporal demand of engagements 2 and 3 would decrease. The SMEs concluded that the reduced temporal demand in engagement 2 would give the pilots more time to complete the doppler notch maneuver (rule 38). In engagement 3, the reduced temporal demand was assumed to leave the pilots in a more favourable position when attempting to adhere to the missile launch range (rule 52), the missile launch parameters (rule 53), and the evasive maneuver initiation range (rule 59). Finally, the SMEs concluded that the modification of qualitative rule 21 (communication of tactical status in engagement 1) had the greatest potential to improve the scores of rules 2 (altitude from 30 seconds to 1 minute since the simulation start in engagement 1) and 12 (timing of the electronic countermeasures in engagement 1). A new version of qualitative rule 21 was also needed to reassure that the adjusted missile launch range in engagement 1 would not damage the score of rule 7. The revision of qualitative rules 44 and 66 (communication of tactical status in engagements 2 and 3) was expected to have a positive impact on the score of rule 47 (altitude from 5 minutes to 5 minutes 45 seconds since

the simulation start) and rule 59 (range of the evasive maneuver initiation) in engagement 3. It was also assumed that the new verbal descriptions of qualitative rules 21, 44 and 66 would have potential to improve SA level 3, as well as to decrease MWL.

In summary, the results of V1 indicated that before proceeding to the L-simulation, it would be beneficial to undertake further TTP refinement in the C- and V-simulations. The main objective of the second C-stage was to modify quantitative rule 7. Qualitative rules 21, 44 and 66 were also under consideration in the second V-stage.

2.4 Second C-stage (C2)

Based on the analysis of V1 results, it was decided to ease rule 7's value (missile launch range in engagement 1). To find a new MP optimal value for rule 7, the original optimization criterion of maximising P_k was relaxed in order to allow a lower P_k compared to the optimal value 0.73 obtained in C1. By following the guidelines discussed in Section 1, the C2 optimization criterion was formulated such that P_{kref}=0.70. The constraint P_s=1.0 was maintained unchanged.

The solution of C2 revealed that the missile launch range in engagement 1 (the value of rule 7) was increased by 17.0%. This provided approximately 11.2 seconds more time for engagement 2. The reduced temporal demand was expected to promote better NP in rule 38 (duration of the doppler notch maneuver in engagement 2), rule 52 (missile launch range in engagement 3), rule 53 (missile launch parameters in engagement 3) and rule 59 (range of the evasive maneuver initiation in engagement 3), as well as to provide lower MWL and higher SA in V2. The adjusted MP optimal rule resulted in P_s=1.0 and an average P_k of 0.70. In engagement 1, P_k was 0.75 and in engagements 2 and 3 P_k was 0.69 and 0.65, respectively.

2.5 Second V-stage (V2)

The original MP optimal quantitative rules with the new MP optimal value of rule 7 (missile launch range in engagement 1) and the original qualitative rules originating from V1 alongside with the revised verbal descriptions of qualitative rules 21, 44 and 66 (communication of tactical status) were applied in V2.

To avoid any improvements in P_k, P_s, NP, SA or MWL because of practice effects, a new group of 14 combat ready male F/A-18C pilots were recruited as participants. The pilots' average

flying experience in F/A-18C was 630 flight hours (SD=334). Based on t-test, there was no statistically significant difference in F/A-18C flight experience between the participants of V1 and V2. V2 was conducted in a same fashion as V1. To simplify the comparison of V1 and V2, the selected results of both stages are next presented.

2.5.1 V2 NP Results

The same rules as in V1 were used in V2 for the NP measure. Based on Mann-Whitney U test, the overall mean of the NP scores in V2 (Mean=0.91, SD=0.29) was significantly higher than in V1 (Mean=0.76, SD=0.42) (U=18.50, p<0.001). In V2, also the NP scores of each engagement were higher than in V1. Table 6 summarizes the descriptive statistics of the NP scores across the engagements in V1 and V2. As in V1, Friedman test revealed significant differences between the V2 NP scores in engagement 1 ($\chi^2_{(21)}$=89.87, p<0.001), in engagement 2 ($\chi^2_{(22)}$=54.96, p<0.001) and in engagement 3 ($\chi^2_{(21)}$=56.98, p<0.001). Figure 5 presents the mean NP scores of V1 and V2 across the TTP rules in engagements 1-3. The associated test statistics and p-values are included in the figure captions.

The rule revision conducted after V1 was mostly successful: while the NP score of rule 46 was slightly lower in V2, the NP scores of all other rules expected to improve from V1, improved in V2. Most of these improvements were statistically significant. According to Figure 5, the NP scores of individual rules in V2 were generally higher or similar to V1. Sixty-five out of 67 NP scores were the same or higher in V2 and the difference of 13 NP scores was significant. Only the scores of rules 29 and 46 were slightly lower in V2, but these differences were not significant.

Table 6. Means and standard deviations (SD) of the NP scores across the engagements in V1 and V2 (N=14).

	V1		V2	
	Mean	SD	Mean	SD
Engagement 1	0.82	0.36	0.92	0.25
Engagement 2	0.76	0.41	0.90	0.23
Engagement 3	0.70	0.39	0.90	0.25

[insert Figure 5]

2.5.2 V2 SA Results

The same SA probes were used in V2 as in V1. Compared to the overall mean SA score of V1 (Mean=0.81, SD=0.39), the overall mean SA score of V2 was higher (Mean=0.89, SD=0.32), but the difference was not statistically significant according to Mann-Whitney U-test. Table 7 summarizes the descriptive statistics of the mean SA scores and the mean SA level scores across the engagements in V1 and V2. In engagements 2 and 3, the mean SA scores in V2 were higher than in V1. In engagement 2, the difference was also statistically significant based on Mann-Whitney U-test (U=38.00, p<0.01). In engagement 2, V2 had significantly higher scores in SA levels 2 (U=59.00, p<0.05) and 3 (U=52.50, p<0.05).

Table 7. Means and standard deviations (SD) of the SA and SA level scores across the engagements in V1 and V2 (N=14).

	SA		SA level 1		SA level 2		SA level 3	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
V1 Engagement 1	0.94	0.24	0.94	0.23	0.99	0.12	0.89	0.32
V2 Engagement 1	0.91	0.29	0.97	0.17	0.96	0.20	0.80	0.40
V1 Engagement 2	0.76	0.43	0.86	0.35	0.84	0.37	0.57	0.50
V2 Engagement 2	0.91	0.28	0.97	0.17	0.97	0.17	0.80	0.40
V1 Engagement 3	0.73	0.45	0.87	0.34	0.77	0.42	0.54	0.50
V2 Engagement 3	0.84	0.36	0.96	0.20	0.86	0.35	0.71	0.46

Friedman test revealed significant differences in V2 SA scores in engagement 1 ($\chi^2_{(14)}$=69.09, p<0.001), engagement 2 ($\chi^2_{(14)}$=72.27, p<0.001), and in engagement 3 ($\chi^2_{(14)}$=51.57, p<0.001). Figure 6 illustrates the mean V1 and V2 SA scores of the probes in engagements 1-3. The associated test statistics and p-values are included in the figure captions.

In V2, 32 of the 45 probe specific SA scores were the same or higher than in V1. In three of those SA scores, the difference was significant. Of the seven probes that had a lower score in V2, the only significant score reduction was in probe 15. The rule modification in V2 failed to improve the scores of probes 14, 29 and 44 related to SA level 3, but it was successful in significantly improving the score of probe 27 (see Figure 6). The new verbal description of the qualitative rule related to the communication of tactical status in engagement 1 (rule 21), was likely to improve the score of probe 26 (see Figure 6). This change was also statistically significant.

[insert Figure 6]

2.5.3 V2 MWL Results

The overall mean of the MWL score in V1 (Mean=47.50, SD=14.44) was higher than in V2 (Mean=42.62, SD=12.69), but the difference was not statistically significant based on Mann-Whitney U test. Table 8 presents the descriptive statistics of the MWL dimensions in V1 and V2. The means of all MWL dimensions were slightly lower in V2, but the differences were not statistically significant.

Table 8. Means and standard deviations (SD) of the MWL dimensions in V1 and V2 (N=14).

MWL dimension	Mean		SD	
	V1	V2	V1	V2
Mental	52.14	49.29	20.82	18.17
Physical	41.43	37.14	21.79	15.90
Temporal	60.00	57.14	18.81	19.39
Performance	44.29	38.57	20.65	18.75
Effort	56.43	46.43	25.90	15.50
Frustration	30.71	27.14	20.93	20.54

2.5.4 V2 Conclusions

Both V1 and V2 resulted in the satisfactory HMP output, i.e., P_k=1.0 and P_s=1.0. Overall, V2 provided a better TTP as the combined effect of the lower temporal demand (see Table 8) and the improved SA level 3 in engagement 2 and 3 (see Table 7) contributed to the overall improvement of NP scores – while HMP output still remained acceptable. A few NP and SA scores remained unchanged or decreased in V2 which suggests that TTP might be further improved by another iteration of C- and V-simulations. However, it was considered that such an exercise would not add value for this demonstration. It was decided to settle for the achieved levels of NP, SA, MWL and HMP output. In other words, the V2 TTP rules were considered as the HMP optimal qualitative and quantitative rules in the simulated environment (see Figure 1). To complete the demonstration of the L-V-C simulation framework, these rules were then evaluated at L-stage using L-simulations to ensure the flight's primary goal is achieved in the light of P_k and P_s, while NP, SA and MWL in the real-life environment remain acceptable.

2.6. First L-Stage (L1)

The HMP optimal rules obtained in V2 were evaluated in L1. Operational pilots and F/A-18C aircraft were used at L-stage. The participants flew as wingmen of a flight and were tasked to follow the HMP optimal rules originating from V2. Supporting pilots flew the other friendly aircraft and were tasked to follow the same rules as the constructive entities did in V2. Pilots of the enemy aircraft were briefed to follow the same scenario used in C- and V-stages.

The participants were given a standard flight briefing, but they were not told how the scenario was to unfold. After the simulation, HMP output, measured by P_k and P_s, and the participants' scores of NP, SA and MWL were observed. Due to a limited availability of F/A-18Cs and pilots, the sample size of L1 was two. These participants did not fly the scenario in V1 or V2 and their average flying experience with F/A-18C was just over 200 hours.

2.6.1 L1 Results

For the most part, the L1 NP scores reflected those of V2. As shown in Figure 7, 56 of the 67 NP scores in L1 were the same or higher than in V2. In L1, both participants failed to adhere to rules 12, 34, and 57 (timing of the electronic countermeasures activation in engagements 1, 2 and 3). It is worth noting that the NP scores of rules 12 and 57 were among the lowest in V2 as well.

[insert Figure 7]

As shown in Figure 8, the L1 SA scores for the probes reflected those obtained in V2. In L1, 38 of the 45 probes had similar or higher scores than in V2. In L1, both participants had difficulties generating alternative courses of action in case something unexpected happened (probes 14 and 44). The scores from probes 14 and 44 were also among the lowest in V2.

[insert Figure 8]

Much as expected, the mean scores on every MWL dimension in L1 were higher than in V2 (see Table 9). It should also be noted that with the exception of the two MWL dimensions with the lowest scores (physical demand, frustration), the order of the dimensions was the same in V2 and L1.

Table 9. Means and standard deviations (SD) of the MWL dimensions in V2 and L1.

MWL	Mean		SD	
Dimension	V2	L1	V2	L1
Mental	49.29	75.00	18.17	7.07
Physical	37.14	40.00	15.90	14.14
Temporal	57.14	85.00	19.39	7.07
Performance	38.57	70.00	18.75	28.28
Effort	46.43	70.00	15.50	14.14
Frustration	27.14	55.00	20.54	21.21

2.6.2 L1 Conclusions

Like V2, L1 resulted in a satisfactory HMP output, as P_k and P_s were both 1.0. While the sample size of L1 did not warrant statistical comparison between V2 and L1, the NP, SA and MWL scores of L1 generally reflected those of V2.

Much like in V2, the timing of electronic countermeasures activation (rules 12, 34, and 57) remained to be an issue in L1. Overall, NP scores were similar and acceptable in V2 and L1 (see Figure 7). As indicated by the SA scores from probes 14, 28, 29 and 44 (see Figure 8), SA level 3 continued to be a challenge in every engagement of L1. In general, the L1 SA scores were acceptable and reflected those of V2. As the real-life task demands and risks motivate the pilots' additional investment of voluntary effort, it was no surprise that all MWL dimensions in L1 were higher than in V2 (see Table 9).

In summary, the evaluation of TTP in L1 implied that the HMP optimal TTP rules (see Figure 1) were applicable in a real-life environment as well. If TTP was to be subjected to further T&E, an additional V-stage (V3) would probably be the most cost-efficient way to address the rules with low scores in V2 and L1. However, as V2 already resulted in the acceptable level of NP, SA, MWL and the decent HMP output in the simulated environment, and L1 strongly suggested that they were acceptable in a real-life environment as well, the demonstration of the L-V-C simulation framework was deemed to be complete.

3. DISCUSSION

This paper demonstrated the L-V-C simulation framework in TTP T&E, where the flight's initial TTP was developed into operational TTP with acceptable HMP output, NP, SA and MWL. In the demonstration, MP optimal quantitative rules were obtained at C-stage, HMP optimal qualitative rules were developed from V-stage, and HMP optimal qualitative and quantitative

rules were evaluated in L-stage using F/A-18C aircraft and qualified fighter pilots. Based on the output of V-stage, TTP rules were identified as candidates for revision in the subsequent C- and V-stages, and the output of V-stage was utilized in the formulation of the follow-on C-stage's optimization criterion. If desired, the results of L-stage could have been used in a same manner in further V- and C-stage iterations.

The demonstration substantiated the usability of the proposed NP, SA and MWL measures in V- and L-stages, as they assisted in identifying candidate rules for revision. The SA scores from V1 and V2 showed how the SA measure was able to identify the increasing SA demands in engagement 3 (see Figure 6 and Table 7), which was mentioned in the initialisation of the example TTP T&E in Section 2.1. As the NP scores of V1 and V2 indicated, NP was a measure with high utility. When combined with the SME analysis, it effectively identified the rules which, once adjusted, improved HMI as demonstrated by improved NP, SA and MWL, and thus avoided premature introduction of L-stage. Finally, the MWL measure also provided complementary results. MWL decreased from V1 to V2, but increased in L1. Overall, all the measures were suitable for the data collection in a natural task setting; their face validity was high, pilots did not report any intrusion, and it was possible to collect the data during normal debriefs.

As the demonstration in this paper highlighted, the measurement of MWL and SA can be of great assistance when interpreting and improving HMI. Whereas P_k, P_s, and NP are strict and well-established measures of goal achievement and task adherence in air combat, existing SA and MWL measures provide more options to choose from. These measures have their individual strengths and weaknesses[4-5], and the T&E setting and objectives should drive the selection of the most appropriate measures. For example, if continuous MWL measurement is a requirement, physiological MWL measures (see, e.g.,[6-9]) should be considered instead of NASA-TLX. Similarly, if the pilot population is not trained to analyze their recollections from the scenario during a debrief, post-trial SA measures may be unreliable.

The use of multiple measures complementing analysis of the outcome of an air combat scenario and pilots' decision making reduces the likelihood of false conclusions about TTP's operational suitability. The diversity of simulation classes and measures lessens also the need that all conclusions drawn from simulation data should be statistically significant. The L-V-C simulation framework should be seen as a decision support tool for the SMEs who ultimately

make the decisions about the rule modifications and TTP's operational approval. While both parts of this paper focused on air combat, the principles of the L-V-C simulation framework are domain independent. As long as there are suitable C-, V- and L-simulation models, the same methodology can be applied to any civil or military task where HMI is of concern.

REFERENCES

1. Mansikka H, Virtanen K, Harris D and Salomäki J. Live-virtual-constructive simulation framework for testing and evaluation of air combat tactics, techniques and procedures, part 1: Theoretical framework. JDMS (submitted for publication).
2. DOD dictionary of military and associated terms, https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/dictionary.pdf (2019, accessed 8 May 2019)
3. Hart SG and Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*. 1988; 52: 139-83.
4. Salmon P, Stanton N, Walker G, Jenkins D, Ladva D, Rafferty L and Young M. Measuring situation awareness in complex systems: comparison of measures study. Int J Ind Ergon 2009; 39: 490-500.
5. O'Donnell RD, Eggemeier FT and Thomas F. Workload assessment methodology. In: K.R. Boff LK and Thomas JP, (eds.). Handbook of Perception and Human Performance. John Wiley and Sons, Inc, 1986, p. 42:1-9.
6. Mansikka H, Simola P, Virtanen K, Harris D and Oksama L. Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. Ergonomics. 2016: 1-9.
7. Mansikka H, Virtanen K, Harris D and Simola P. Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. Appl Ergon 2016
8. Roscoe A. Assessing pilot workload. Why measure heart rate, HRV and respiration? Biol Psychol 1992; 34: 259-287.
9. Jorna P. Heart rate and workload variations in actual and simulated flight. Ergonomics 1993; 36: 1043-1054.

FIGURE 1

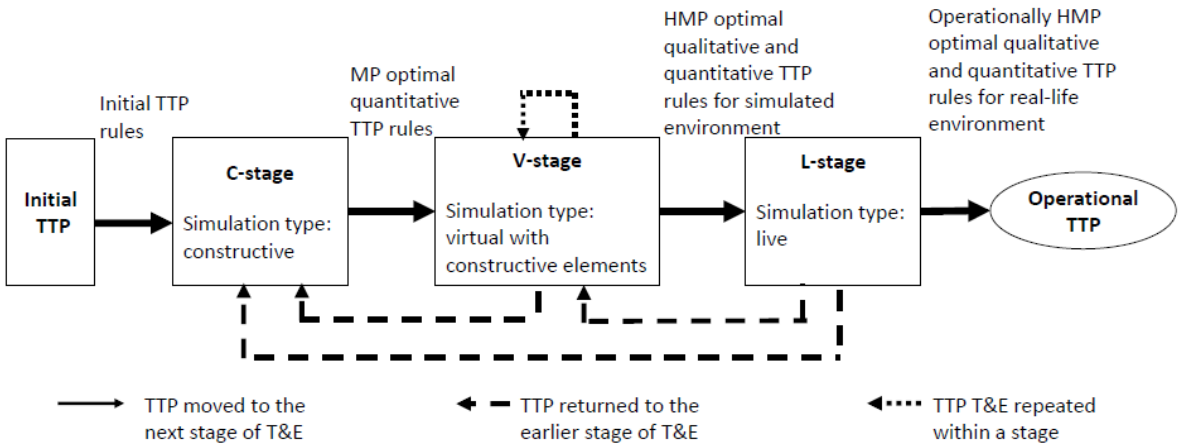
Initial TTP rules
MP optimal quantitative TTP rules
HMP optimal qualitative and quantitative TTP rules for simulated environment
Operationally HMP optimal qualitative and quantitative TTP rules for real-life environment
Initial TTP
C-stage
Simulation type: constructive
V-stage
Simulation type: virtual with constructive elements
L-stage
Simulation type: live
Operational TTP
TTP moved to the next stage of T&E
TTP returned to the earlier stage of T&E
TTP T&E repeated within a stage

FIGURE 2

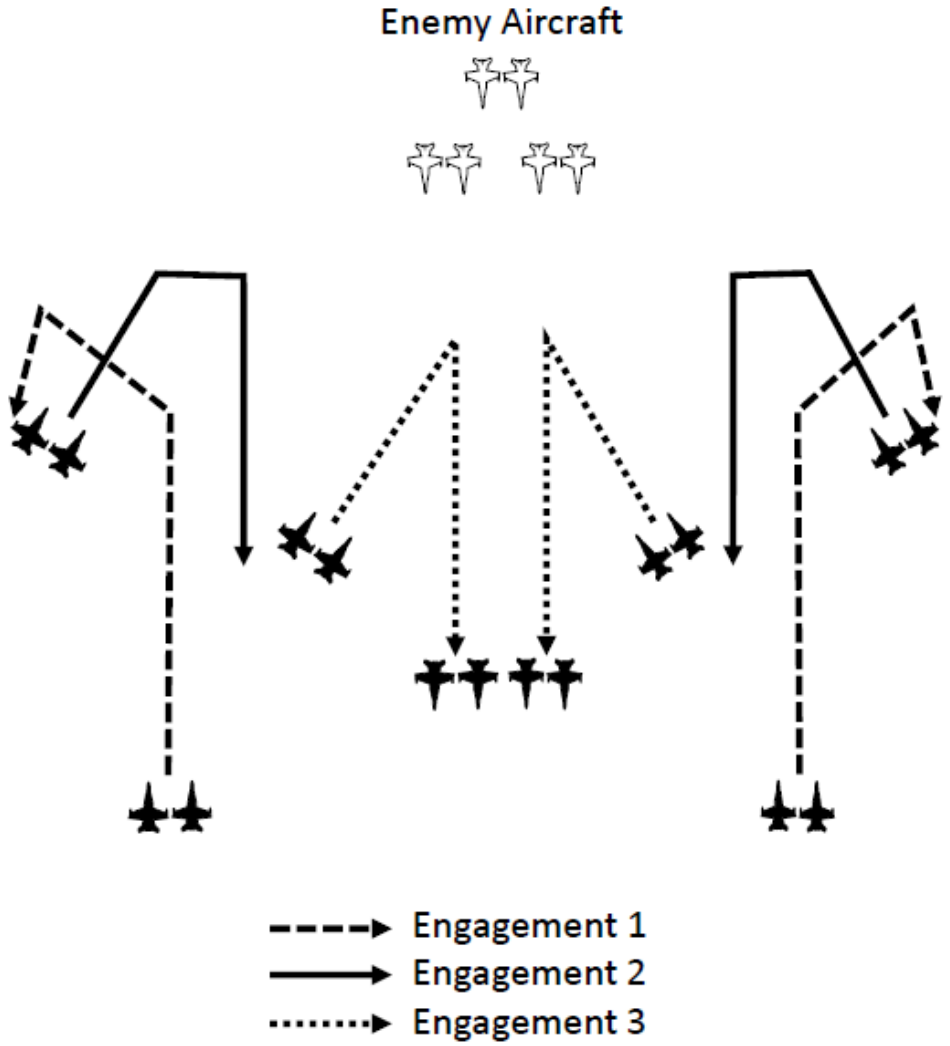

FIGURE 3

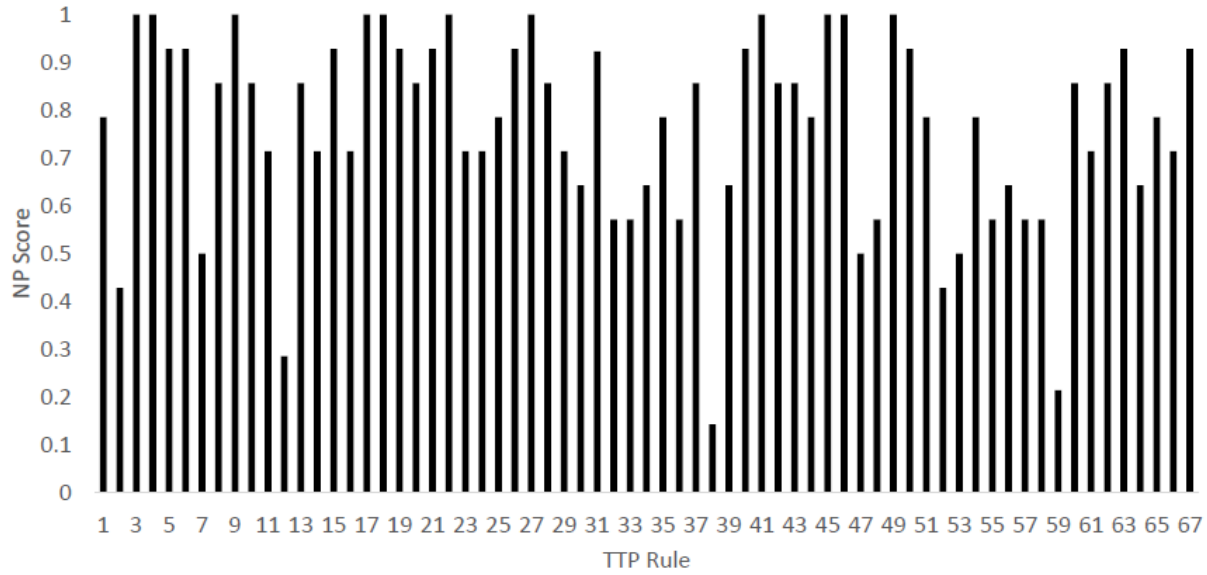
NP Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67
TTP Rule

FIGURE 4

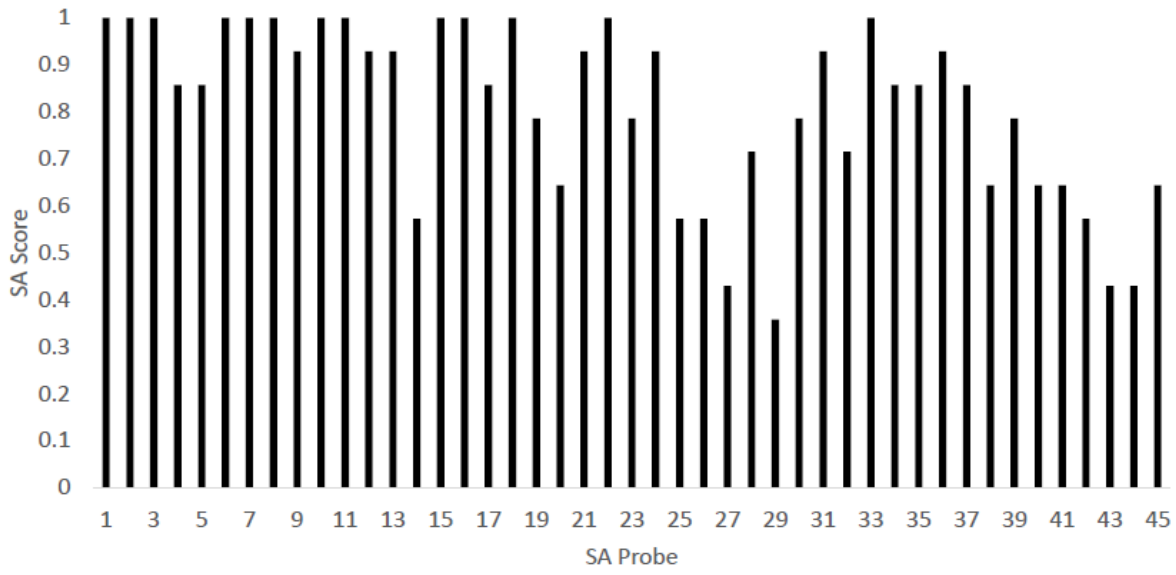
SA Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
SA Probe

FIGURE 5

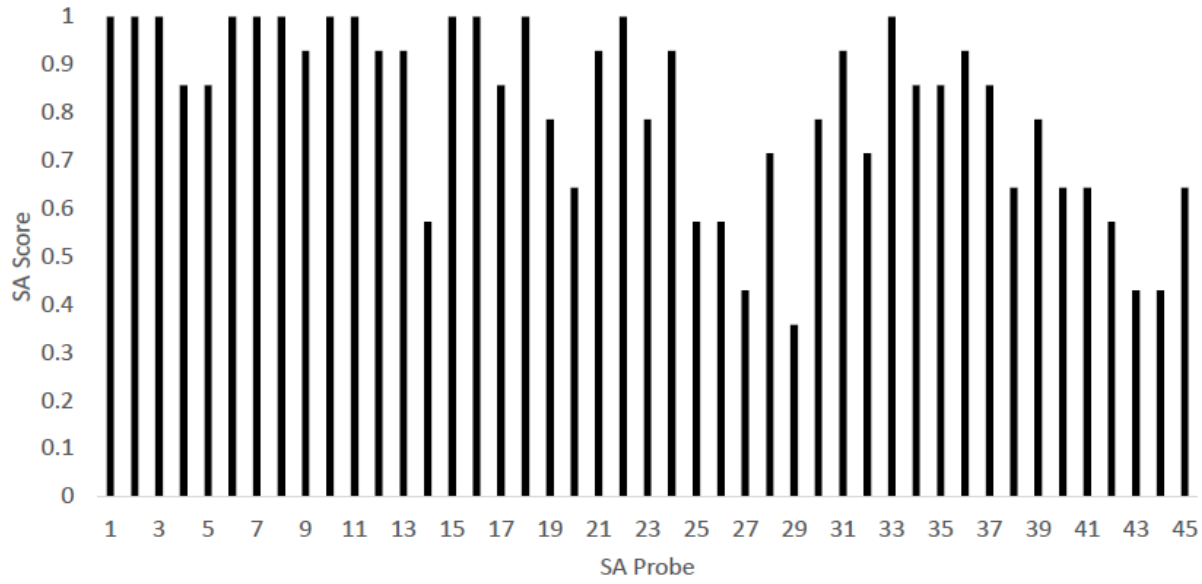
SA Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
SA Probe

FIGURE 6

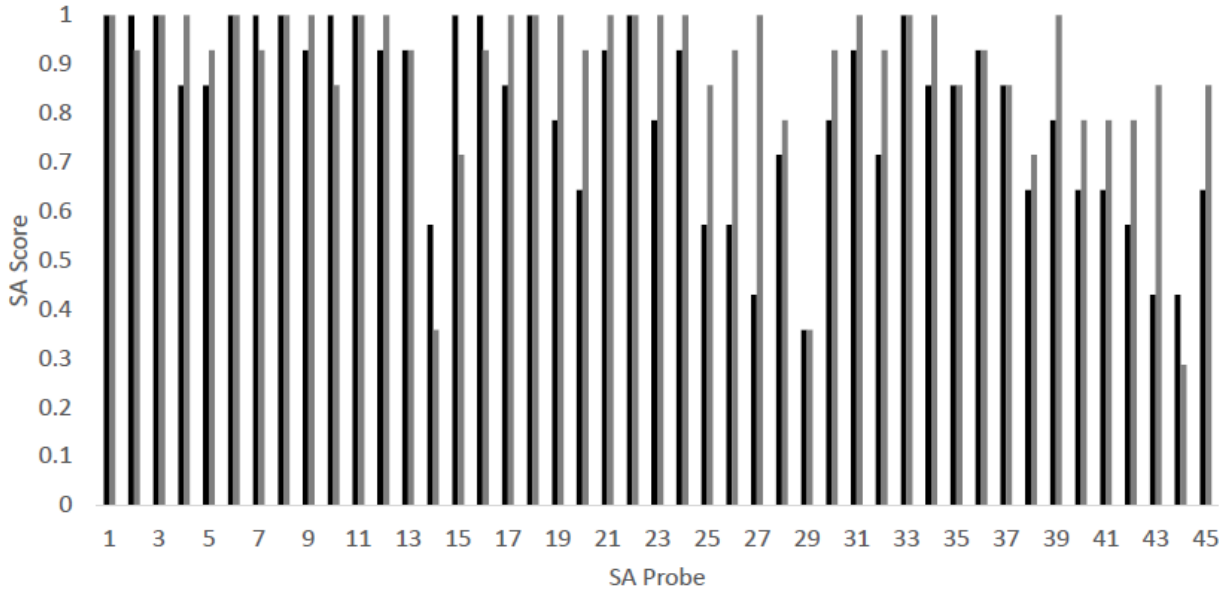
SA Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
SA Probe

FIGURE 7

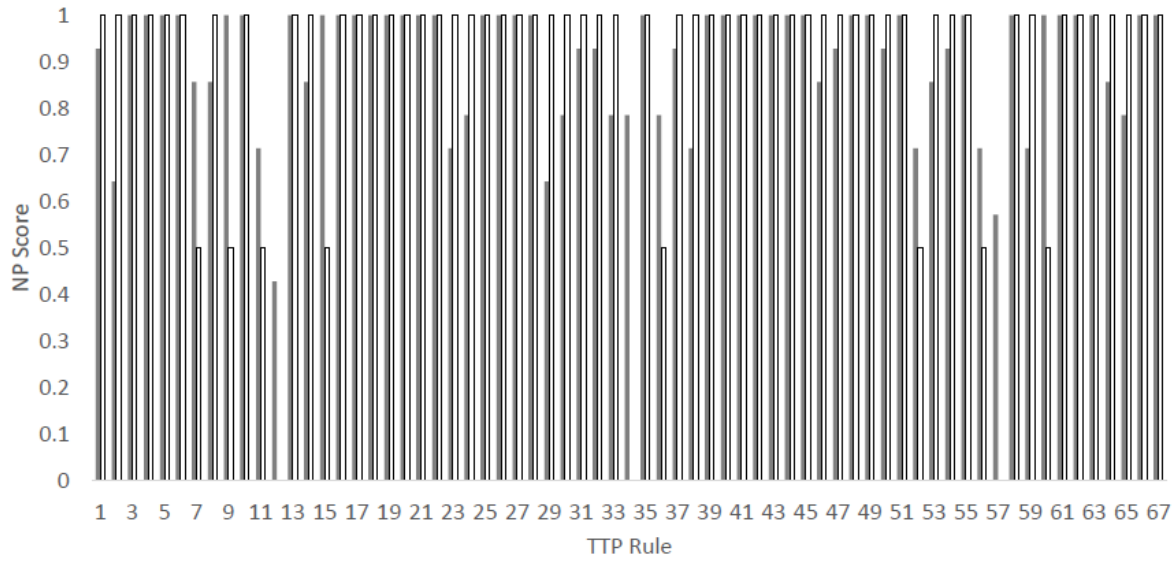
NP Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67
TTP Rule

FIGURE 8

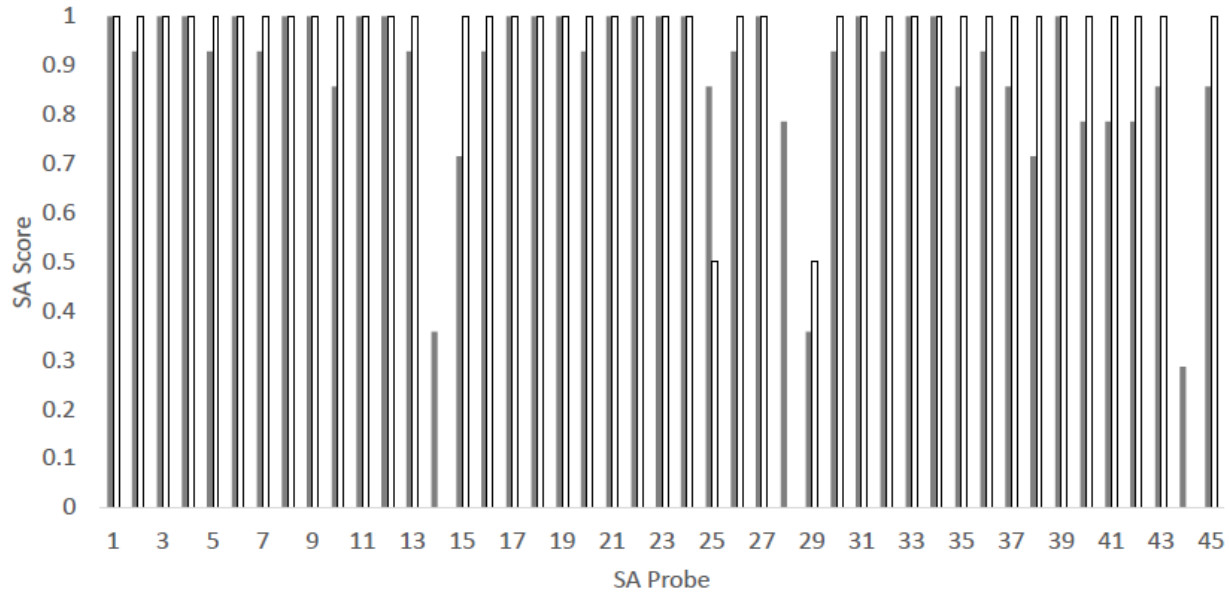
SA Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
SA Probe

Figure 1. Structure and operating principle of the L-V-C simulation framework.

Figure 2. Scenario and the overall flow of the BVR air combat task.

Figure 3. Mean NP scores across TTP rules in engagements 1-3 (see Table 1 for descriptions of rules).

Figure 4. Mean SA scores across probes in engagements 1-3 (see Table 3 for descriptions of probes).

Figure 5. Mean NP scores across TTP rules in engagements 1-3. Black bars denote V1 and grey bars V2, respectively. In engagement 1, rules 7 and 21 were modified for V2. An improvement in the mean NP score was expected in rules 2, 7, 12, and 21. Based on Mann-Whitney U-test statistically significant differences were observed in rule 7 ($U=63.00$, $p<0.05$) and rule 16 ($U=70.00$, $p<0.05$). In engagement 2, rule 44 was modified for V2. An improvement in the mean NP score was expected in rules 38 and 44. Statistically significant differences were observed in rule 32 ($U=63.00$, $p<0.05$), rule 38 ($U=42.00$, $p<0.01$), and rule 39 ($U=63.00$, $p<0.05$). In engagement 3, rule 66 was modified for V2. An improvement in the mean NP score was expected in rules 47, 52, 53, 59 and 66. Statistically significant differences were observed in rule 47 ($U=56.00$, $p<0.05$), rule 48 ($U=56.00$, $p<0.05$), rule 53 ($U=63.00$, $p<0.05$), rule 55 ($U=56.00$, $p<0.05$), rule 58 ($U=56.00$, $p<0.05$), rule 59 ($U=49.00$, $p<0.01$), rule 61 ($U=70.00$, $p<0.05$), and rule 66 ($U=70.00$, $p<0.05$).

Figure 6. Mean SA scores across probes in engagements 1-3. Black bars denote V1 and grey bars V2, respectively. Probes 11-15 were associated with SA level 3, probes 26-30 were associated with SA level 3 and probes 41-45 were associated with SA level 3. Based on Mann-Whitney U-test a statistically significant difference was observed in probe 15 (did you correctly anticipate the result of the engagement?) ($U=70.00$, $p<0.05$), probe 26 (where you able to correctly anticipate how the engagement evolved?) ($U=63.00$, $p<0.05$), probe 27 (were you able to correctly anticipate the actions of your flight members?) ($U=42.00$, $p<0.01$) and probe 43 (were you able to correctly anticipate the ranges and decision points in the timeline?) ($U=56.00$, $p<0.05$).

Figure 7. Mean NP scores across TTP rules in engagements 1-3. Grey bars denote V2 and white bars L1, respectively.

Figure 8. Mean SA scores across probes in engagements 1-3. Grey bars denote V2 and white bars L1, respectively.